State of California

Office of the State Chief Information Officer

Geospatial Information

Standards Document

December 2009

Document History

Revision History			
Revision	**Date of Release**	**Owner**	**Summary of Changes**
Initial Release (v 0.1)	December 2009	Michael Byrne	Initial Release – describing Geocoding

Table of Contents

Contents

Geospatial Information Standards
Executive Summary

The Office of the State Chief Information Officer (OCIO) for the State of California provides leadership for the State's information technology (IT) programs and works collaboratively with other information technology leaders throughout state government. The OCIO's role, therefore, is as a strategic planner and architect for the State's information technology programs and as a leader in formulating and advancing a vision for that program.

There is a growing demand for the State of California to conduct its business differently. California has a significant challenge to redesign its business approaches and processes. Its greatest challenge is to implement an IT environment that supports a business model that builds an infrastructure which connects agencies to each other and their customers and provides appropriate access to information from any place, at any time. This new business model includes: (1) coordinated service delivery across agencies; (2) citizen–centric one stop shopping; (3) more planned and coordinated partnerships with external organizations; and (4) streamlined administrative business processes.

Geographic Information Systems (GIS) are critical to these goals and if used effectively, can help facilitate the changes required with the state. It is a core strategy of the State to translate its data to spatial data (e.g. map based data), use geographic approaches in analyzing its data and publish data (where appropriate) as spatial data. Described here are standards, procedures, and protocols which establish the statewide roadmap to achieve the business mission and goals by improving the performance of its core business processes through effective GIS implementation and geospatial data management.

1.0 Introduction

Geographic Information Systems (GIS) are systems of hardware and software interacted with by users which store/retrieve, analyze, display, capture and output spatial data for the intent of making decisions. The State of California has a long history of GIS use in several departments. Some departments have moved or are moving towards enterprise implementation of GIS solutions. While some industry standards are predominantly adopted, this document is the first to describe State of California standards, protocols and procedures for GIS. Chapters will be developed over time. This document is intended to be a living document, governed by the OCIO, which articulates the standards, protocols and procedures for implementing GIS solutions in the State.

1.1 Purpose

The OCIO prepared this document to assist Agencies and departments in (1) developing a common implementation for GIS across the state, (2) providing a roadmap for those agencies not yet implementing GIS solutions and (3) developing procedures for governing GIS data. This document will be edited by the OCIO as technology changes and advances.

1.2 Background

As described in Government Code Section 11545, the Office of the State Chief Information Officer (OCIO) has responsibility for guiding the application of IT in California State government. This includes establishing and enforcing state IT strategic plans, policies, standards and enterprise architecture. Key areas described in the IT Capital plan include the use of Geographic Information Systems. As such the OCIO is taking an industry standards approach to GIS, and in particular outlining standards which will (1) foster more effective and efficient government through common operating procedures (e.g. GIS procedures which affect the most number of department's business processes will be outlined first) and (2) build common infrastructure for maintaining state data. Standards development will be a best of breed as described by, among others, the Federal Geographic Data Committee, Federal Enterprise Architecture, the Open Geospatial Consortium and industry standards.

1.3 Related Activities

The following lists the current initiatives and how they interrelate to the GIS data efforts defined within. **IT Capital Plan –** The IT Capital Plan collects the Administration's plan for strategic IT investments. This plan provides the strategic direction required for developing EA's TO-BE targets that are used to support development of planned investments. **Data Strategy Initiative –** The purpose of Data Strategy initiative is for the OCIO to gain an understanding of the systems and data that exists and how the data is used across the enterprise of the State of California.

2.0 Geocoding

2.1 Introduction

Geographic data exists in State databases and files that describe locations such as street addresses, city names, ZIP Codes, and even telephone numbers. While people understand what these descriptions mean and how they relate to locations on the earth's surface, analysis for the intent of decision making in a computer can do little with this information. Geocoding is the process of assigning an X,Y coordinate value to a place by comparing the descriptive location elements to those present in a reference data set. These X,Y coordinate pairs are point locations which can be displayed on a digital map and prepared for use in analysis.

Once the location information is transformed into a common two-dimensional coordinate system (X and Y) various analytical techniques can be performed that can disclose patterns and relationships that are not apparent in tabular data. Patterns that can be analyzed include but are not limited to;

Issue	Description
Fraud	Potential fraud cases because the normal distribution of government funds (e.g. unemployment insurance or public health assistance) is far exceeded at a given address
Health Patterns	Identification of health concerns, like clusters of a disease, outside the normal rate of infection and hospitalization issues from a set of addresses within a neighborhood
Environmental Patterns	Evaluation of environmental concerns given the concentration of toxic release issues within a range of addresses
Taxation	Assessment of fair tax collection and levees given the normal distribution of taxation at a set of addresses
Economic Development	Strategic planning of business development and growth in the state given addresses of employer locations and access to business tax incentives
Government Services	Identification for the need for government services given the types of populations at a set of addresses
Government Services	Evaluation of the effectiveness, equity and transparency of government services at a set of addresses

The above issues, and countless other government business objectives cannot be met unless the State transforms its descriptive location data to an X,Y coordinate system. Geocoding is the process by which this is accomplished.

Data that is geocoded has a common element for database management; the X,Y. These common elements can be used to help populate and effectively manage ancillary data often required for the state's data systems, because the X,Y is a unique identifier for place. This unique identifier can be used to populate the ancillary data. This ancillary data can be elements like the county name, legislative district, and special district. Moreover, geocoding can standardize the descriptive data so that there is a single collection of real identified addresses, street names, cities and ZIP codes. If proper geocoding techniques are used, fields like these

can be automatically populated reducing human entry time and error caused by human data entry. A proper geocoding infrastructure can help more effectively manage our State databases.

2.2 Definitions

Below are definitions pertinent to the geocoding process.

Geocode – A standardized representation for a location given a textual description of the location like an address, ZIP Code, or place name. The standardized representation is typically an X,Y coordinate and/or a latitude and longitude. Generally speaking, geocode refers only to a street address text description of place.

Composite Service – a service used to provide a geocode address whereby multiple layers of address data are used in a hierarchy to achieve maximum accuracy.

CASS – The United States Postal Service Coding Accuracy Support System. The CASS enables the USPS to evaluate the accuracy of address-matching software programs and provide grades for vendors of software. In addition, the vendors then have an ability to change and modify their software to increase the accuracy of address-matching functions. Many currently available software packages meet the CASS standard.

2.3 Descriptions

Geocoding has three main components; Reference Data, Address Data and software. These components are described here. In addition composite geocoding is described.

Reference Data

Reference data is the geographic base files on which a geocoding engine runs. These base files are most often address point locations, street center line data, ZIP Code boundaries and place name locations. Reference data is available from both public and private sources. Public data sources include the US Census Bureau TIGER Line files, and the US Postal Service city/state five digit ZIP codes and ZIP+4 files. Numerous private sources provide licenses to reference data depending on business need.

The reference data must be stored and maintained in the standard projection and datum. Establishing this allows for the geocoding service to return the standardized projection for the X,Y coordinate. A projection algorithm can be used to return the Latitude and Longitude of the X,Y in the standard datum (See Projection Standard).

Address (or Source) Data

Address or Source data is the descriptive place data owned by the business application. These data vary by type, but in terms of State government it is most often a physical street address

(1325 J Street). Examples of Departments keeping street address data as the corporate enterprise data source include, among others, the Franchise Tax Board, the Board of Equalization, the Secretary of State, the Employment Development Department, the Department of Health Care Services, the Department of Public Health and the Department of Social Services. In addition, nearly every department in state government has a data set of street address locations for managing its own building location services and outlets for public access (e.g. Department of Motor Vehicle office locations). Additionally Source Data can be incident or event locations (e.g. a fire, or vehicle accident), location of equipment and facilities (e.g. medical stockpile locations, hospitals), and monument locations (e.g. .1 miles north from intersection of Hwy 49 and Interstate 80 on Hwy 49).

Much of Source Data is human entered and can contain errors in address formatting or transposed elements (e.g. numbers entered incorrectly or transposed, misspelled addresses, not real addresses and non-standardized address elements (e.g. Boulevard vs. BLVD). These errors, however, can be corrected through a standardization process at the onset of geocoding.

Software (or geocoding application)

Software, or the geocoding application, can perform many functions. The basic function is to parse the Source Data into defined elements for better understanding and matching. In a physical address example this parsing includes the address number, street prefix, street name, the street suffix, street direction, street type (e.g. BLVD), City name, ZIP Code, ZIP+4, and State. Software then performs a probabilistic record linkage with a statistically valid form of fuzzy logic to score how well the Source Data can be matched to the Reference Data. This type of matching allows for reviews of “almost” matches, scoring thresholds, index tuning, best match and/or candidate matching.

Geocoding software can also perform a standardization process, whereby the Source Data is standardized along say the US Postal Service standards. This process increases the match rate potential of the Source Data and provides for a common storage taxonomy for the Source Data. Geocoding software can return, depending on the solution architecture, an X,Y location, the standardized address, a score, a match sequence (e.g. what reference data it matched to), and ancillary data as required.

It is important to distinguish between software designed to geocode for the purpose of performing spatial analysis and software designed to geocode for the purpose of minimizing the expenses related to undeliverable and duplicate mail. Some mail standardization software companies advertize that the software can also geocode. However, this software generally uses less reliable street segment data, and the unmatched rates and number of records assigned to the ZIP Code centroid would be higher, and not adequate for spatial analysis.

2.4 Standards

The following standards are approved for State of California geocoding methods. This standard is a ‘process’ standard. The description below is the appropriate steps to follow to comply with the state standard for managing address data.

Step 1: Define Input or Source Data – Field Definitions

It is the standard of the State of California for agencies to store a minimum of the following fields for descriptive location (eg address) data. It is acceptable to modify the field definitions to meet specific business needs (e.g. change field name and/or lengths). However, it is not acceptable to not carry one of the below fields along with the geocoded record.

Field	Type	Width	Description
First Address Field	Text	50	Street number, and name, intersections acceptable. Example: 1325 J Street
Second Address Field	Text	40	Building, floor, mail stop, PO Box, suite, etc. Example: Suite 1600
City	Text	30	City Name Example: Sacramento
State	Text	2	State Abbreviation Example: CA
ZIP Code	Text	5	ZIP Code Example: 95814
ZIP 4 Extension (optional)	Text	10	ZIP Code + "-" and 4 digit extension if available Example: 95814-1234.

Step 2 – Standardization and Validation Process

Once data is assembled in the proper format, it is the standard of the State of California for geocoding processes to use a United States Postal Service Coding Accuracy Support System Certified (http://www.usps.gov/cass.htm) software as the standard to validate and standardize street address information prior to geocoding. This software can be part of a single geocoding package, a stand alone software and/or a web service.

Step 3: Composite Geocoding Process

A composite geocoding service will rely on multiple reference data, and in order of best quality attempt to match the Source Data in a cascading fashion. If no match is reached in the best quality reference data set, the service will turn to the next reference data set and attempt to perform the same function and repeat until a match is found which meets the business rules applied to the reference data. A composite geocoding service performs the same function as a single reference geocoding service plus the reference layer it matched to. In this environment record level analysis can then determine the rate for best quality versus lower quality matches in addition to finding some location for nearly every single Source Data record rather than just leaving them Unmatched. The State of California standard geocoding application is a composite service to ensure the most possible location matches.

Step 4 – Keep Post Geocoded Address Fields

It is the standard of the State of California for agencies to keep a minimum of the following fields from the results of the geocoding process. It is acceptable to modify the field definitions to meet specific business needs (e.g. change field name and/or lengths). However, it is not acceptable to delete any of the below fields along with the geocoded record. A typical industry geocoding software will return all of these fields.

Field	Type	Width	Description
Geocoding Score	Integer	3	A score rating the quality of address match (geocode) to a reference data set. Business rules for minimum match score are the responsibility of the sponsoring agency for defining. Example: 100
Geocoding reference data	Text	50	A string indicating the source of the reference data to which the address was matched Example: 1-TA_Points_ZIP_0708
Standard Address	Text	50	The standardized and validated address captured for address data quality and management (output of Address 1). Example: 2575 Sand Hill Rd
Standard City Name	Text	30	The standardized and validated city name captured for data quality and management (output of City). Example: Davis
Standard ZIP Code	Text	9	The standardized and validated ZIP +4 captured for data quality and management (Output of ZIP) Example: 95814-0000
Longitude	Floating decimal	6 Decimal	The x-coordinate of the geocoded address in geographic projection (NAD83 – see projection standard). A minimum of 6 decimals must be carried in this field, depending on the business need. Example: -120.554987
Latitude	Floating decimal	6	The y-coordinate of the geocoded address in geographic projection (NAD83 – see projection standard). A minimum of 6 decimals must be carried in this field, depending on the business need. Example: 37.491958
X	Floating decimal	3	The x-coordinate of the geocoded address in standard California Albers Projection (NAD83 – see projection standard). A minimum of 3 decimal places must be carried in this field.

Field	Type	Width	Description
			Example: -100125.1234
Y	Floating decimal	3	The y-coordinate of the geocoded address in standard California Albers Projection (NAD83 – see projection standard). A minimum of 3 decimal places must be carried in this field. Example: 43678.1234

Step 5 – Final Data Base Management

Agencies are able to maintain these fields in data models suiting their appropriate needs, as long as a minimum number of the above fields are maintained at the record level.

2.5 Best Practices

The below listed best practices are the recommended procedures for geocoding. Agencies and Departments should make their best efforts to follow these guidelines.

Business Needs

It is a State of California best practice that the business unit develops the business specifications and need for analysis of the data prior to geocoding. Having an understanding of how geocoded information will benefit the business are prior to geocoding will benefit both the information technology and business units mutually.

Composite Service

It is a State of California best practice to develop and use composite geocoding services in the following order:

- Address Point or site address parcel centroid then;
- Street center line file (best possible industry licensed data if funding permits)
- Street center line (publically available – US Census TIGER)
- ZIP+4 centroids
- ZIP (5 digit only) centroids
- Place name (e.g. city name)

Address Entry

It is a State of California best practice to geocode the data at the time of data entry, so that address standardization, validation and end user confirmation of the correct location can occur. While we understand that address geocoding often cannot happen at the time of entry and that batch geocoding is often the only opportunity to geocode, the best practice is to implement system design criteria such that geocoding CAN happen at the time of entry. In this instance

the data entry person can be given the opportunity to edit and correct errors thus eliminating significant amounts of wasted time and effort later in the process.

Offset

It is a State of California best practice, when using street centerline data, to off-set the resulting geocode to the appropriate street side as defined by the "Left Side" or "Right Side" in reference center line data by a minimum of 15 meters. Offsetting geocodes improves the reliability of ancillary data management returned from the geocoding process (e.g. assigning a county, legislative or other district to the geocode).

Non-Matched Records

It is a State of California best practice to assess, attempt to identify locations for non-matched records and store any non-matched records. It is the goal of the state to have the highest quality data as possible. Even in the best scenarios 95% - 99% record match is commonly extremely high. What this illustrates is that any geocoding activity means there will be non-matched records. It is the duty of the data managers and owners to perform the following;

- Assess these non-matched records for simple transposition mistakes in the addresses and fix where appropriate or the data owners can. Assessment includes documentation of the geocoding base score (e.g. number of matched records) as a quality score of the dataset. In some cases the data volume is too large and the business case does not require 100% geocoding match. In these cases the best practice is for the data owner to make this assessment;
- Identify locations for non-matched records can be accomplished through many software applications. In the simple validated address which cannot be matched might occur due to some error in the address (e.g. transposed number or prefix etc). The best approach is for the data managers make any correction required in the source address, and use the geocoding application to rematch the address. Often the best approach is to hand place the location on a map base. This location approach is dependent on the appropriateness of the business need.
- In the case of non-matched records after an attempt to rematch (described above) the best practice is to store the non-matched records along with all the returned fields even though the match score is 0. Keeping these records as part of the dataset demonstrate the whole dataset and intrinsically identify error in the data.

3.0 Projection

3.1 Introduction

Projection is a mathematical method of representing the surface of a sphere or other shape on a flat two-dimensional plane. A cornerstone of GIS functionality, and mapping in general, is the ability to reference data, which exists on the surface of the earth (a 3 dimensional sphere) in a two dimensional plane (e.g. a paper, graphic or internet map). The mathematical transformations for these processes attempt to preserve one of three qualities, while distorting the other two. These qualities are shape, area and direction. For instance, some projections tend to hold area as constant as possible, so land area can be consistently measured over the

curvature of the earth in the case of a large variation in degrees north and south, like California. In this case, shape and direction are distorted more than area. Projections are developed to take into account not only these qualities, but the proximity to the polar regions, the size of the study area, the variation over north/south degrees and interest in a metric or English unit (meters vs. feet).

3.2 Definitions

Projection – A method for representing the surface of a sphere or other shape on a flat two-dimensional plane.

Albers Equal Area Projection – A projection which is specifically designed for maintaining area over shape and direction, particularly over large north south variation.

Geodetic Datum – A standard position or level that measurements are taken from. A datum is used to define projections further, in order to deal with the fact that the earth is not a true sphere.

3.3 Standard

Geodetic Datum

The State of California has codified in Public Resouces Code Section 8850 – 8861 (see http://www.leginfo.ca.gov/cgi-bin/waisgate?WAISdocID=0664971606+0+0+0&WAISaction=retrieve) the official geodetic datum and reference network for use within the State. State spatial data are required to be collected and stored in this standard.

Projection

The State of California standard geospatial projection is defined as below. This definition is a common configuration for projection files in industry GIS software.

Albers Equal Area	
Projection	Albers
Units	Meters
1st standard parallel	34 00 00
2nd standard parallel	40 30 00
Central Meridian	-120 00 00
Latitude of Origin	00 00 00
False easting (meters)	0
False northing (meters)	-4,000,000
Datum	NAD83
Spheroid	Clarke 1866

Other projections can be used in analysis, storage and retrieval of GIS system implementations. For the sake of publishing, geocoding and interoperability, the California Albers projection defined above shall be used.

Coordinate System

When storing geographic data as un-projected, the State of California standards is:

Geographic Coordinates	
Datum	NAD83